

A Further (Itakura-Saito/ $\beta = 0$) Bi-stochasticization and Associated Clustering/Regionalization of the 3,107-County 1995-2000 U. S. Migration Network

Paul B. Slater*

*University of California,
Santa Barbara, CA 93106-4030*

(Dated: October 26, 2012)

Abstract

We extend to the β -divergence (Itakura-Saito) case $\beta = 0$, the comparative bi-stochasticization analyses—previously conducted (arXiv:1208.3428) for the (Kullback-Leibler) $\beta = 1$ and (squared-Euclidean) $\beta = 2$ cases—of the 3,107-county 1995-2000 U. S. migration network. A heuristic, “greedy” algorithm—using the $\beta = 1$ results as an initial configuration—is devised. While the largest 25,329 entries of the 735,531 non-zero entries of the bi-stochasticized table—in the $\beta = 1$ case—are required to complete the widely-applied two-stage (double-standardization and strong-component hierarchical clustering) procedure, 105,363 of the 735,531 are needed (reflective of greater uniformity of entries) in the $\beta = 0$ instance. The North Carolina counties of Mecklenburg (Charlotte) and Wake (Raleigh) are considerably relatively more cosmopolitan in the $\beta = 0$ study. The Colorado county of El Paso (Colorado Springs) replaces the Florida Atlantic county of Brevard (the “Space Coast”) as the most cosmopolitan, with Brevard becoming the second-most. Honolulu County splinters away from the other four (still-grouped) Hawaiian counties, becoming the fifth most cosmopolitan county nation-wide. The five counties of Rhode Island remain intact as a regional entity, but the eight counties of Connecticut fragment, leaving only five counties clustered.

PACS numbers: Valid PACS 02.10.Ox, 02.10.Yn, 89.65.Cd, 89.75.Hc

*Electronic address: slater@kitp.ucsb.edu

We continue our comparative investigations of bi-stochasticizations of weighted, directed networks—in particular, the network of 1995-2000 migration flows between 3,107 U. S. counties—and their associated clustering/regionalization properties [1].

We have previously ”bi-stochasticized” the $3,107 \times 3,107$ matrix of flows by minimizing each of two forms of β -divergence. ”The β -divergence is a family of cost functions parameterized by a single shape parameter β that takes the (squared)-Euclidean distance, the Kullback-Leibler divergence and the Itakura-Saito divergence as special cases ($\beta = 2, 1, 0$ respectively)” [2].

We—in the extensive series [3, 4] of applications of the two-stage (double-standardization [5], followed by strong-component hierarchical clustering [6]) algorithm—had always employed the well-established Kullback-Leibler-based procedure ($\beta = 1$) for double-standardization [7]. In [1], we, for the first time, implemented the $\beta = 2$ approach [8, 9], and found strong differences between the $\beta = 2$ and $\beta = 1$ results. In particular, in the $\beta = 2$ case, there were 2,707 entries of the associated doubly-stochastic matrix equal to the (maximum possible value of) 1, while in the doubly-stochastic matrix for $\beta = 1$, there was only a single such entry.

Here, we seek to expand this pair of analyses to also include the $\beta = 0$ (Itakura-Saito) case. Not being aware of any specific effective algorithm for this purpose [9, p. 357], we developed a heuristic ”greedy” procedure. It relies upon the availability (as a starting point) of the previous results of the $\beta = 1$ bi-stochasticization.

We proceed by randomly choosing a pair (m_{ij}, m_{kl}) of the 735,531 non-zero entries in the original data (flow) table. If $i \neq k$ and $j \neq l$, then we ask if m_{il} and m_{kj} are also non-zero. If so (which occurs about 9.22% of the time), we seek that (arbitrarily-signed) value of x which when added to m_{ij} and m_{kl} and subtracted from m_{il} and m_{kj} minimizes the (Burg-entropy-based [11, Table 2.1]) objective function

$$\frac{m_{ij}}{s_{ij} + x} - \log \frac{m_{ij}}{s_{ij} + x} + \frac{m_{kl}}{s_{kl} + x} - \log \frac{m_{kl}}{s_{kl} + x} + \frac{m_{il}}{s_{il} - x} - \log \frac{m_{il}}{s_{il} - x} + \frac{m_{kj}}{s_{kj} - x} - \log \frac{m_{kj}}{s_{kj} - x}, \quad (1)$$

where we impose the constraints, $0 < s_{ij} + x < 1, 0 < s_{kl} + x < 1, 0 < s_{il} - x < 1, 0 < s_{kj} - x < 1$. Here the s ’s, initially, are chosen to be the corresponding entries of the $\beta = 1$ bi-stochasticized table, previously obtained (using the well-known Sinkhorn-Knopp iterative algorithm [7]). Then, the four indicated entries are updated by either adding or subtracting the optimal value of x . This procedure, importantly, preserves the bi-stochasticity of the

$\beta = 1$ bi-stochastic table from which we have started our heuristic, "greedy" procedure.

The initial value of the sum

$$\sum_{i,j}^{n=3107} \left(\frac{m_{ij}}{s_{ij}} - \log \frac{m_{ij}}{s_{ij}} - 1 \right), \quad (2)$$

which is taken (thus, avoiding singularities) only over the 735,531 non-zero entries ($m_{ij} > 0$) of the $3,107 \times 3,107$ table, was 4.71219×10^{11} . Implementing the minimization operation (1) 82 million times, and updating the values of the s 's as we proceed, we reduced this sum to 1.59538×10^{11} . (On the other hand, the objective function—the Kullback-Leibler divergence [that is, $x \log \frac{y}{x} + y - x$]—in the $\beta = 1$ case, achieving a minimum value of 4.92974×10^8 there, increases to 5.01181×10^8 , if the $\beta = 0$ bi-stochastic table, derived from the $\beta = 1$ table as its starting point, is substituted in the objective-function calculation.) To indicate the strong convergence of the algorithm, the objective function after 80 million iterations was 1.59541×10^{11} .

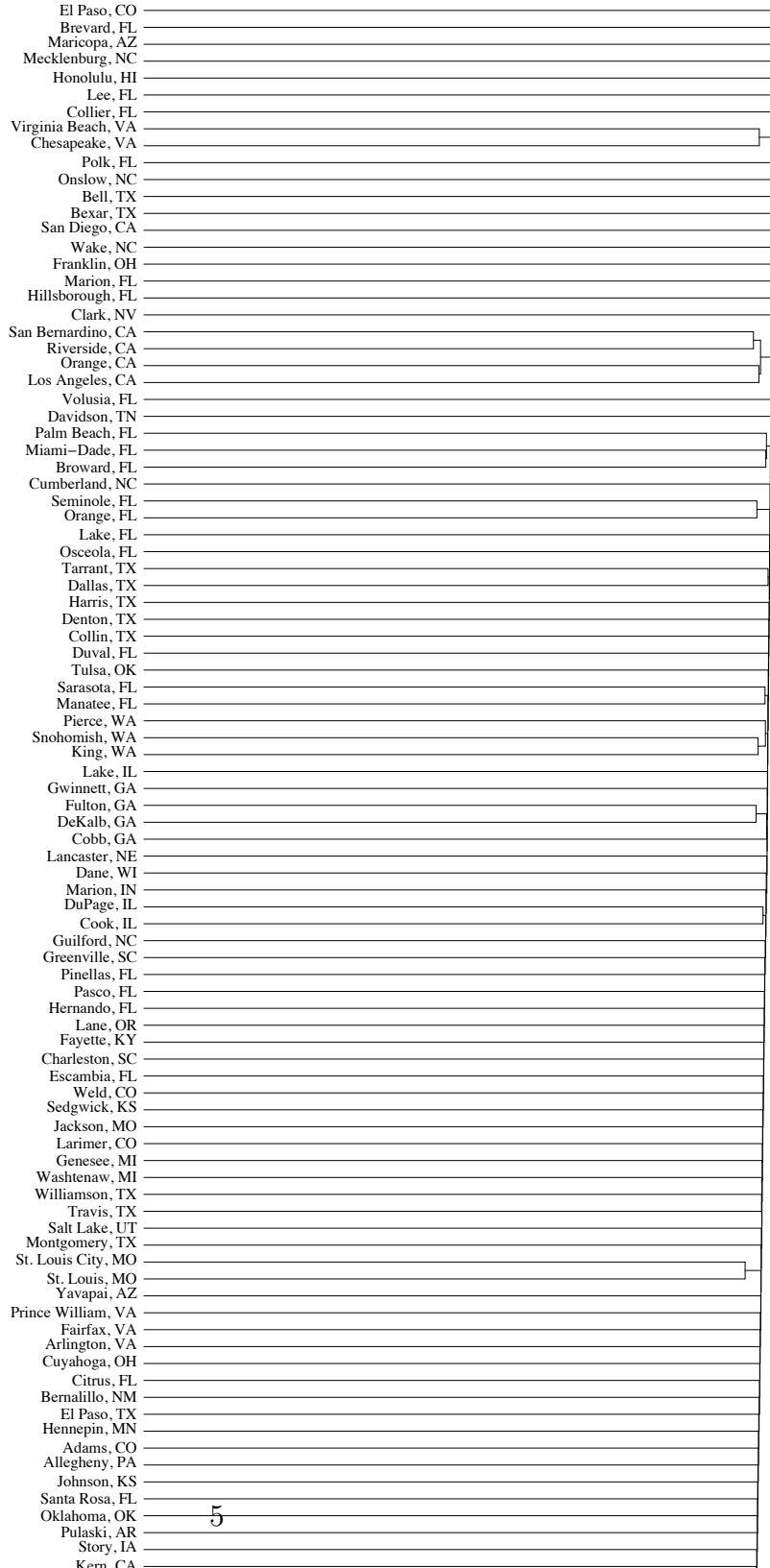
Next, applying the strong-component hierarchical clustering step [6] of the two-stage algorithm [3, 4]—with 2,517 non-trivial mergings occurring (2,497 for $\beta = 1$)—the largest 105,363 entries of the $\beta = 0$ table were required to complete the clustering, while only 25,239 were needed in the $\beta = 1$ case [10]. (This appears to be indicative of the greater uniformity of entries in the $\beta = 1$ analysis. The $\beta = 2$ case, on the other hand, did not seem to lend itself meaningfully to the application of the hierarchical clustering procedure, due to the large concentration [87.1284%] of its non-zero entries equalling 1, as well as its relatively small number [57,153 *vs.* 735,531] of strictly non-zero entries.)

We, now, present the (ordinally-ranked) dendrogram associated with the $\beta = 0$ analysis, while its $\beta = 1$ counterpart can be viewed in [10]. The North Carolina counties of Mecklenburg (Charlotte) and Wake (Raleigh) are considerably relatively more cosmopolitan in the $\beta = 0$ study than in the $\beta = 1$ analysis, as well as Franklin County, Ohio (Columbus, the state capital). The Colorado county of El Paso (Colorado Springs) replaces the Florida Atlantic county of Brevard (the "Space Coast") as the most cosmopolitan, with Brevard becoming the second-most. Only five of the eight counties of Connecticut are clustered in the $\beta = 0$ analysis, while all eight form a well-defined region in the $\beta = 1$ case. The five counties of Rhode Island are grouped in both studies, but the fifth county (Honolulu) of Hawaii is now omitted from the state grouping in the $\beta = 0$ study, becoming the fifth most cosmopolitan nation-wide.

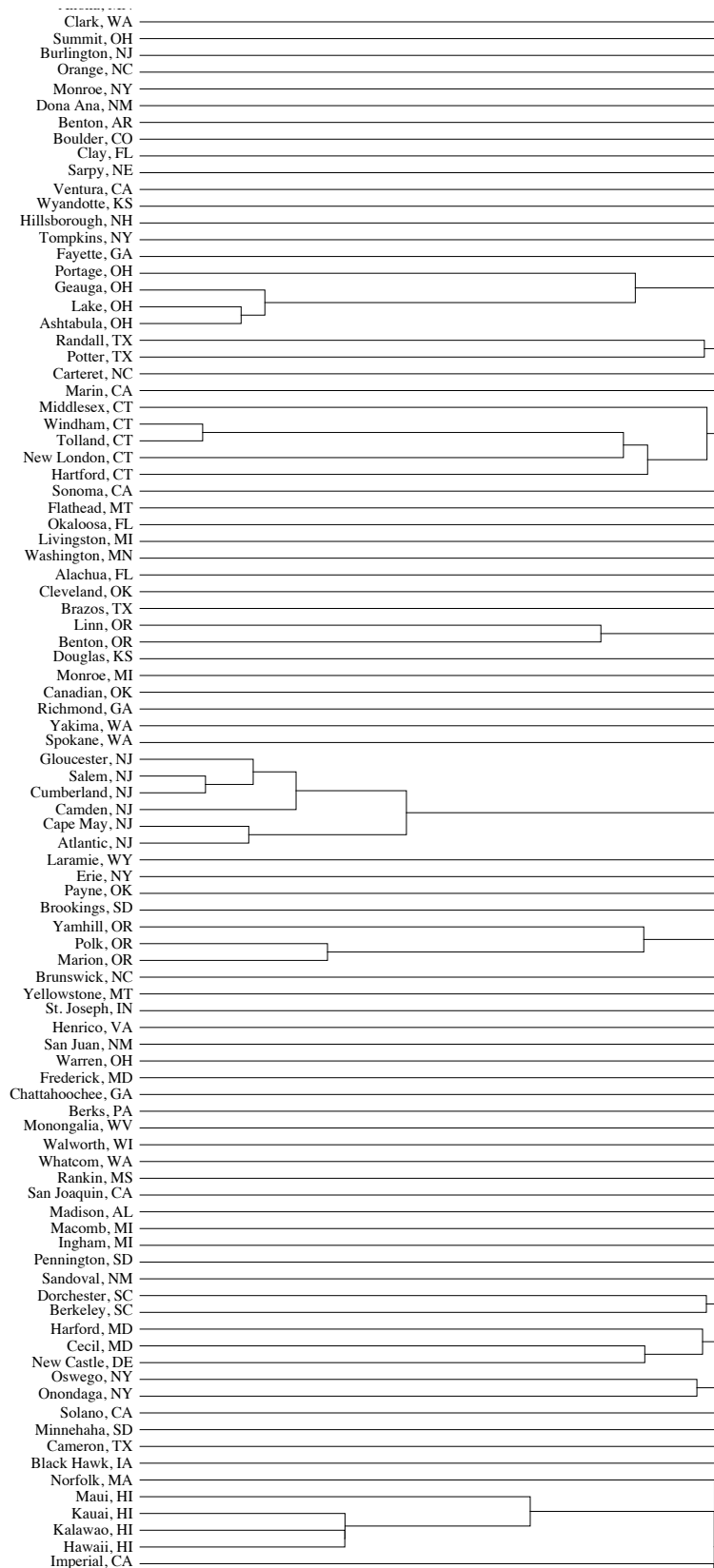
Some "fine-tuning" of our clustering results may be subsequently reported, as we continue to run our algorithm, obtaining ever-increasing degrees of the already high convergence already achieved.

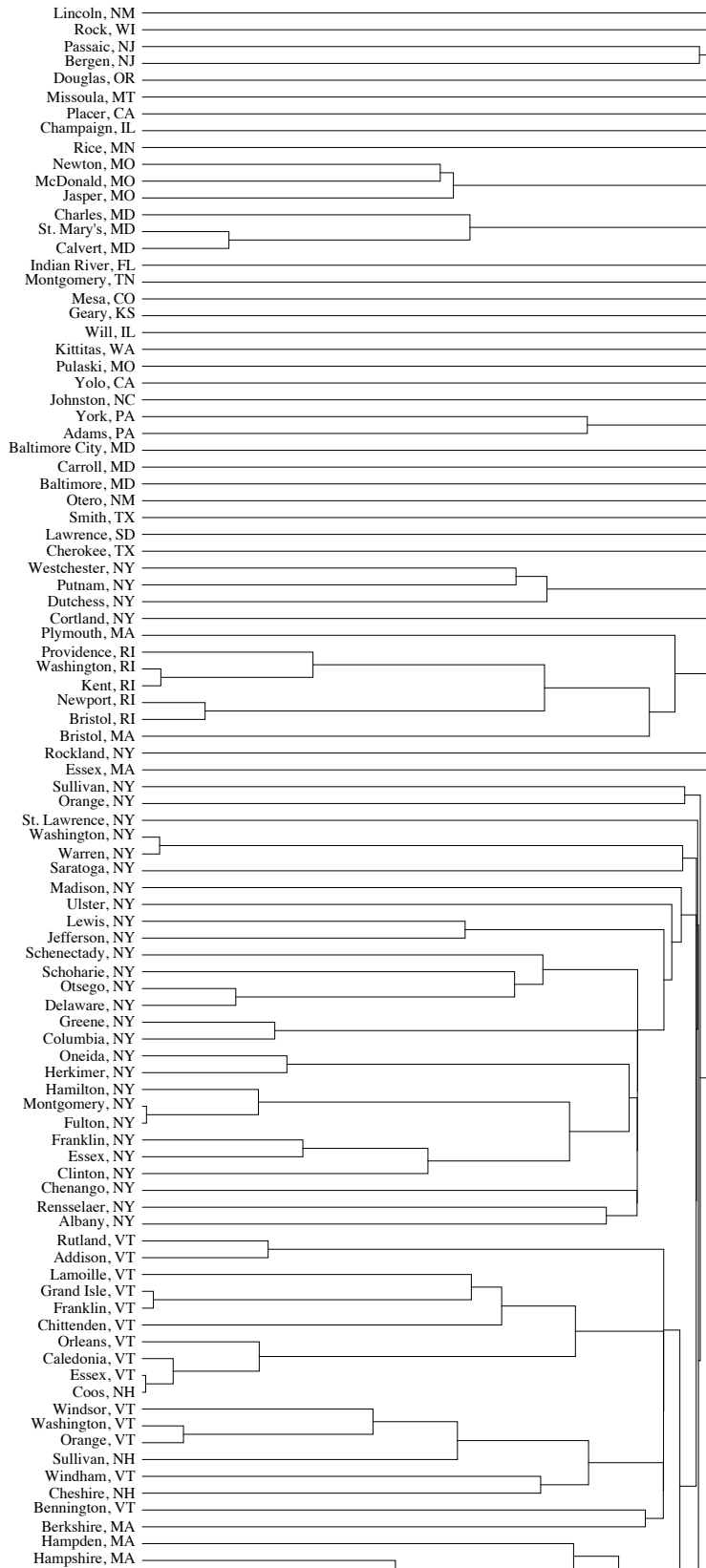
We are also exploring the use of additional forms of Bregman divergences—such as the inverse $(\frac{1}{x})$ type [11, Table 2.1].

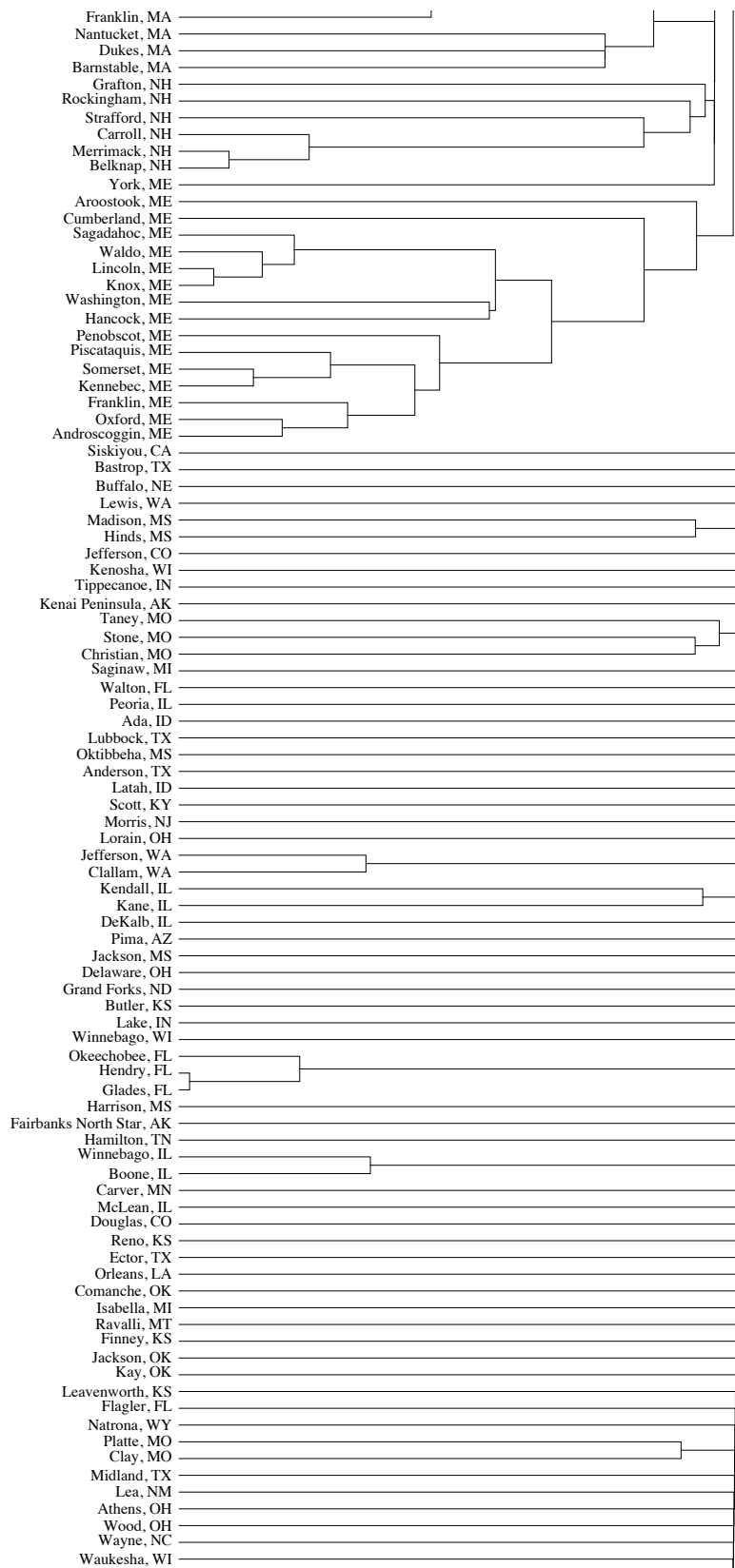
I. COUNTY-LEVEL DENDROGRAM

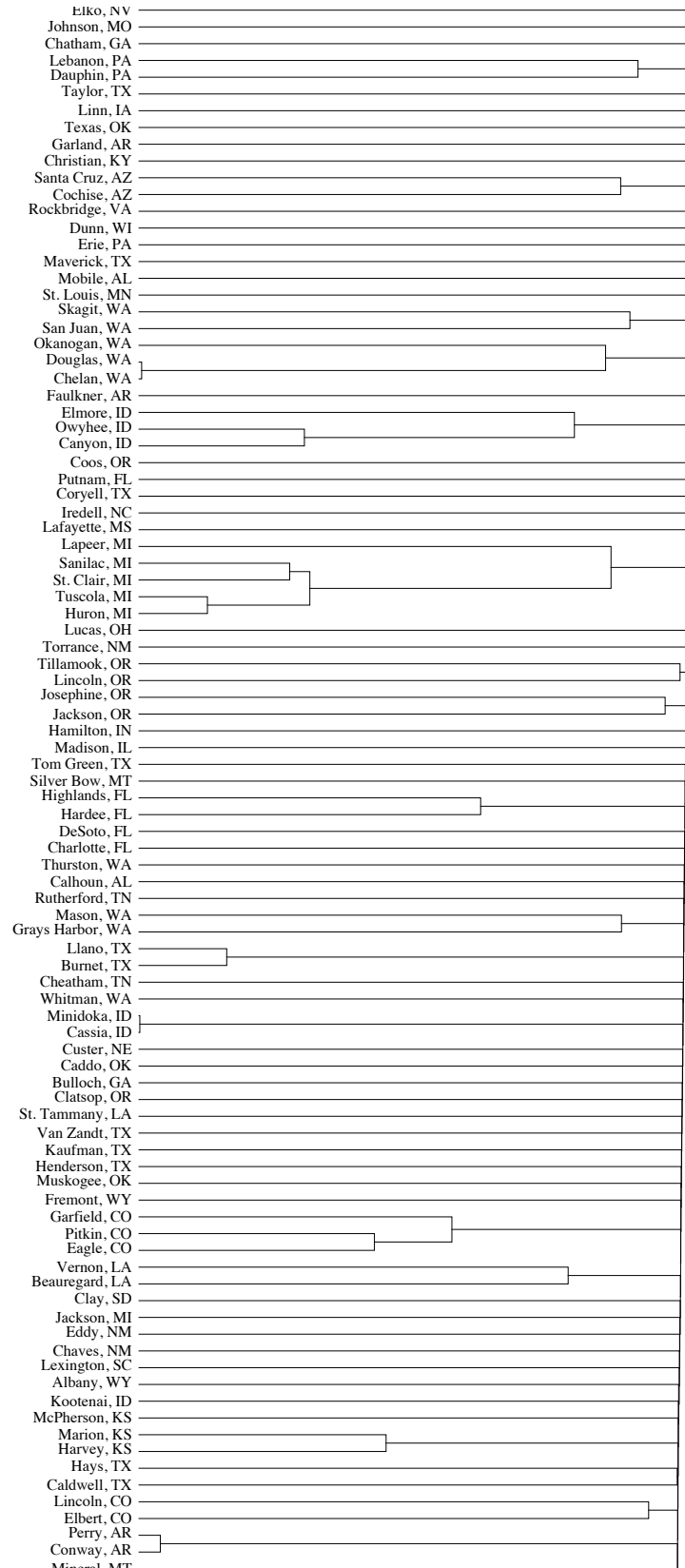


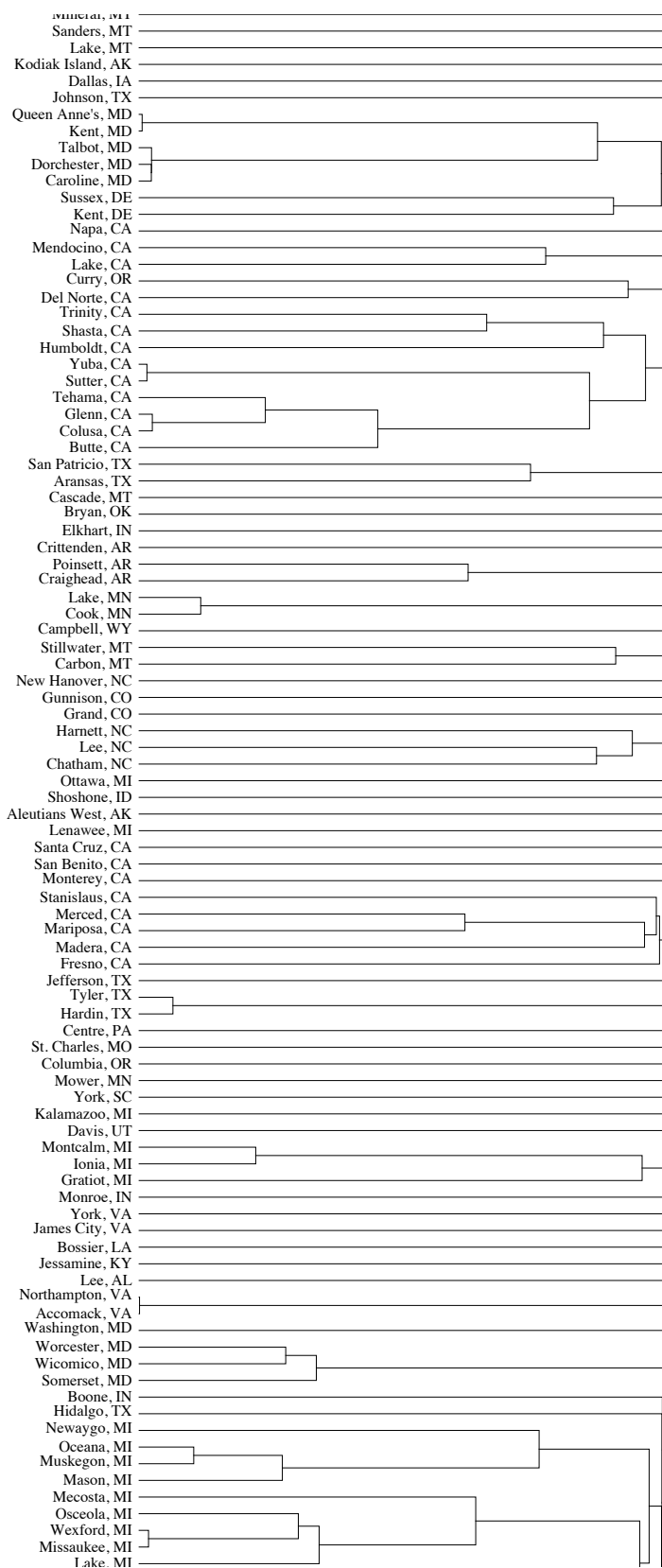
Kenil, CA	
Santa Barbara, CA	
San Luis Obispo, CA	
St. Johns, FL	
Polk, IA	
Nueces, TX	
Brazoria, TX	
Hudson, NJ	
Worcester, MA	
Anne Arundel, MD	
St. Lucie, FL	
Martin, FL	
Galveston, TX	
Montgomery, MD	
Philadelphia, PA	
Delaware, PA	
Chester, PA	
Montgomery, PA	
Bucks, PA	
Wayne, MI	
Pinal, AZ	
Richland, SC	
Washington, OR	
Multnomah, OR	
Clackamas, OR	
Fairfield, CT	
Hamilton, OH	
Newport News, VA	
Hampton, VA	
Jefferson, AL	
Denver, CO	
Mercer, NJ	
Contra Costa, CA	
Alameda, CA	
Montgomery, OH	
Greene, OH	
Mohave, AZ	
Gallatin, MT	
Santa Clara, CA	
San Mateo, CA	
San Francisco, CA	
Ramsey, MN	
Dakota, MN	
Douglas, NE	
Shelby, TN	
Sacramento, CA	
Queens, NY	
Suffolk, NY	
Nassau, NY	
Clayton, GA	
New York, NY	
Kings, NY	
Boone, MO	
Ocean, NJ	
Monmouth, NJ	
Howard, MD	
Horry, SC	
Kent, MI	
Prince George's, MD	
District of Columbia, DC	
Richmond, NY	
Monroe, FL	
Arapahoe, CO	
Sumter, FL	
Durham, NC	
Tulare, CA	
Kings, CA	
Jefferson, KY	
Anchorage, AK	
Lancaster, PA	
Oakland, MI	
New Haven, CT	
Suffolk, MA	
Middlesex, MA	
Fort Bend, TX	
Island, WA	
Riley, KS	
Litchfield, CT	
Milwaukee, WI	
Kitsap, WA	
Johnson, IA	
Baldwin, AL	
McHenry, IL	
Williamson, TN	
Butler, OH	
Cass, ND	
Union, NJ	
Essex, NJ	
Yuma, AZ	
Bronx, NY	
Greene, MO	
Middlesex, NJ	
Scott, IA	
Anoka, MN	

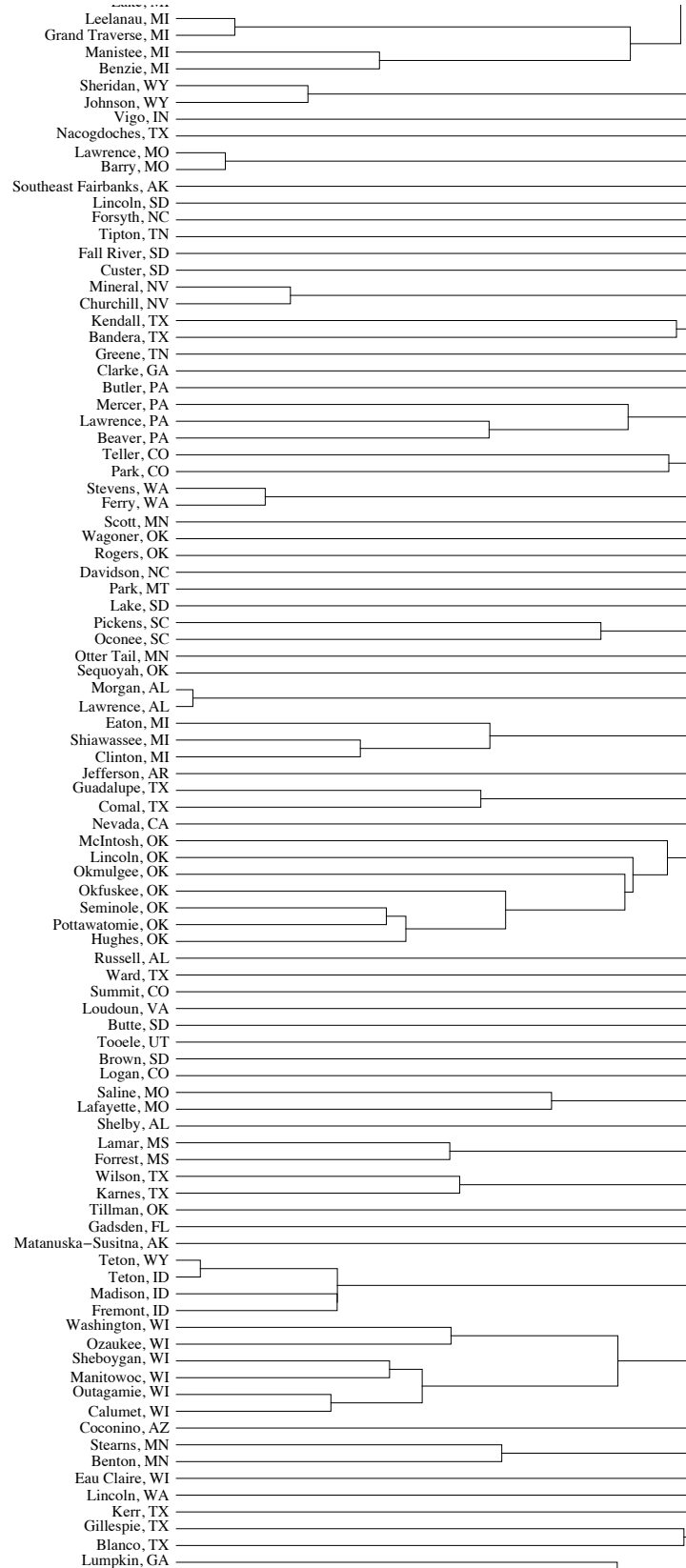


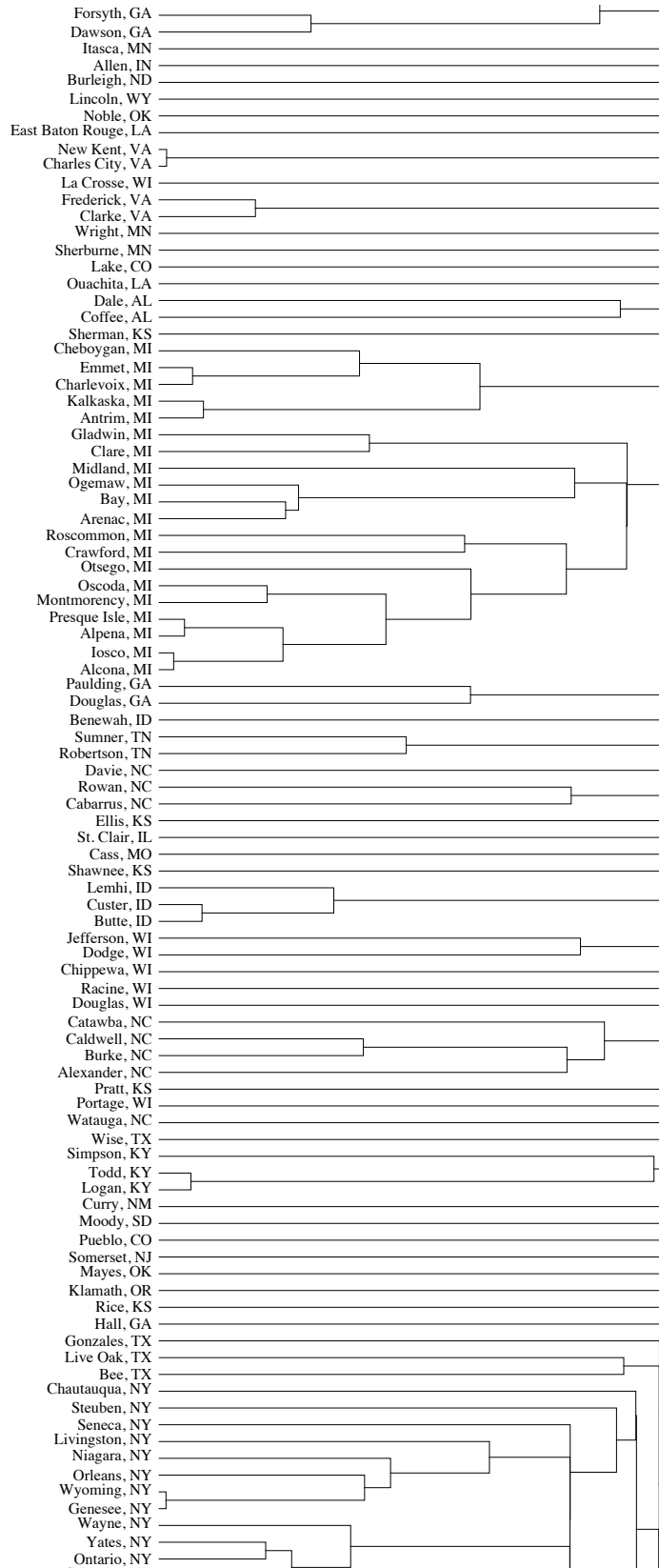


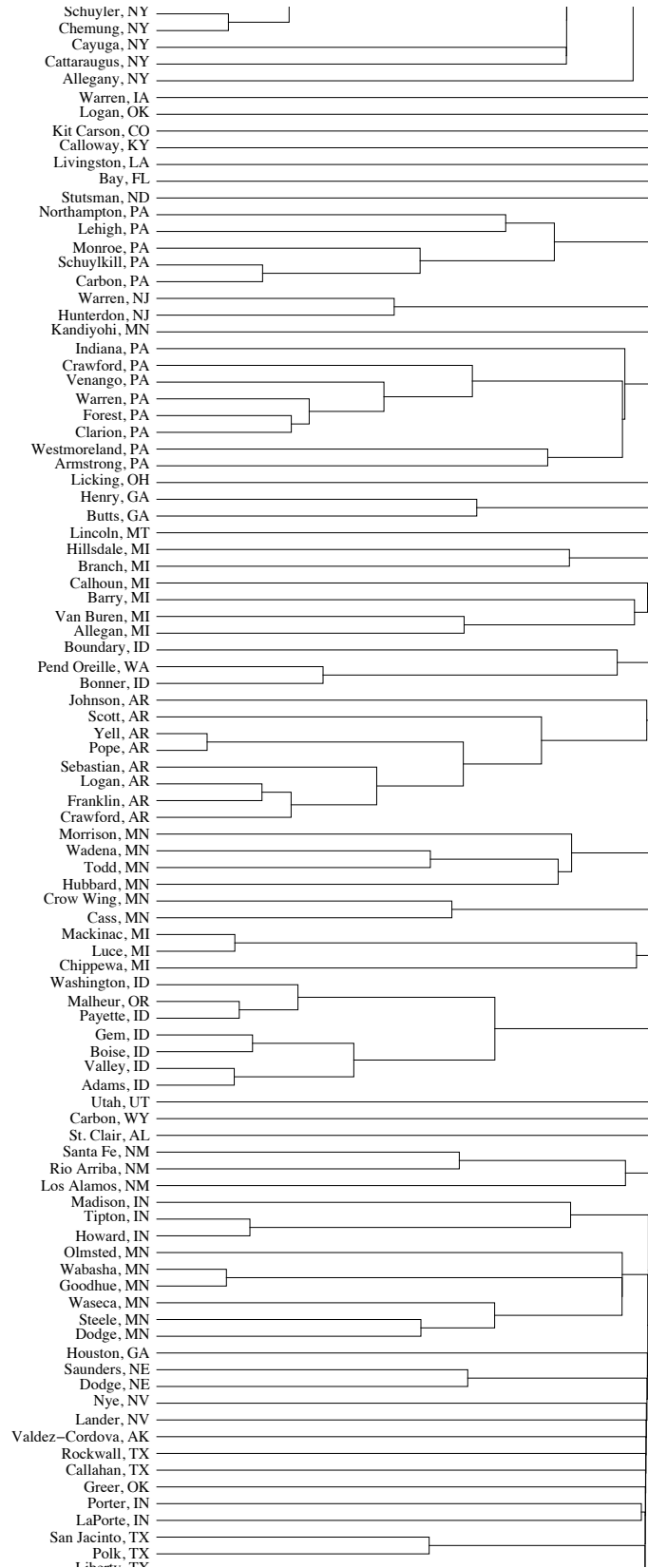


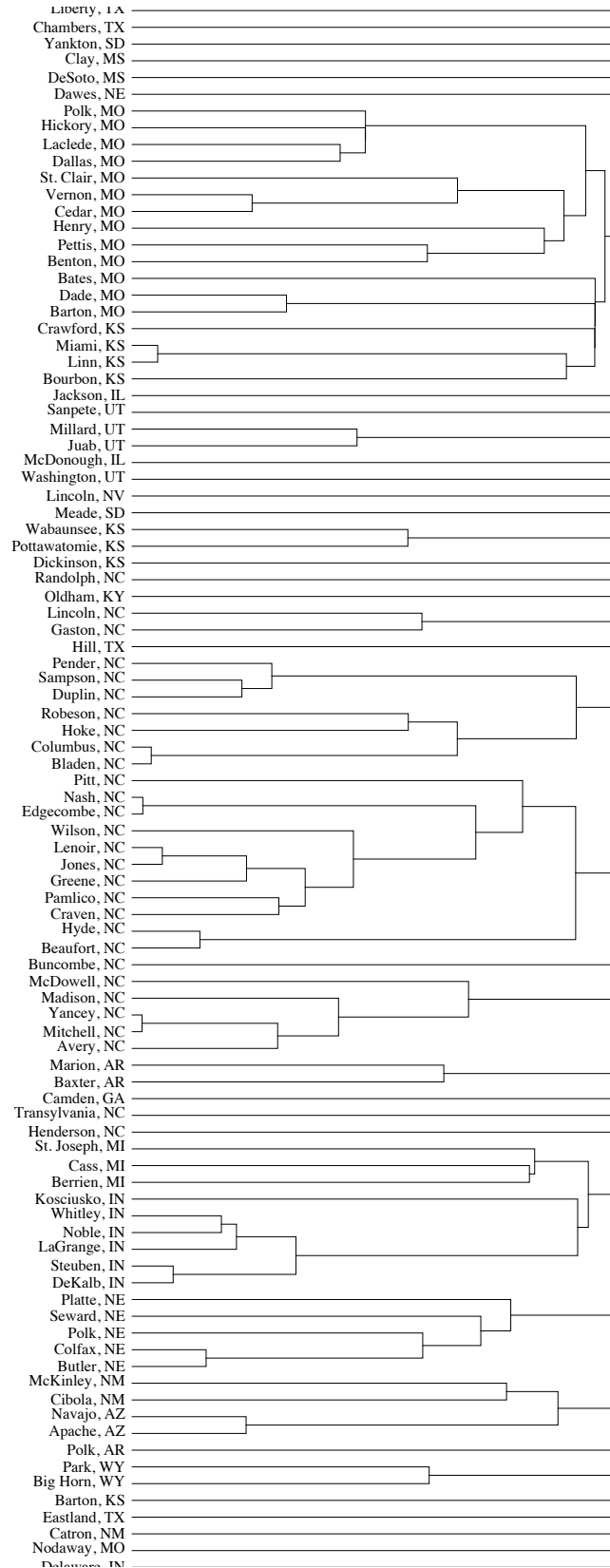


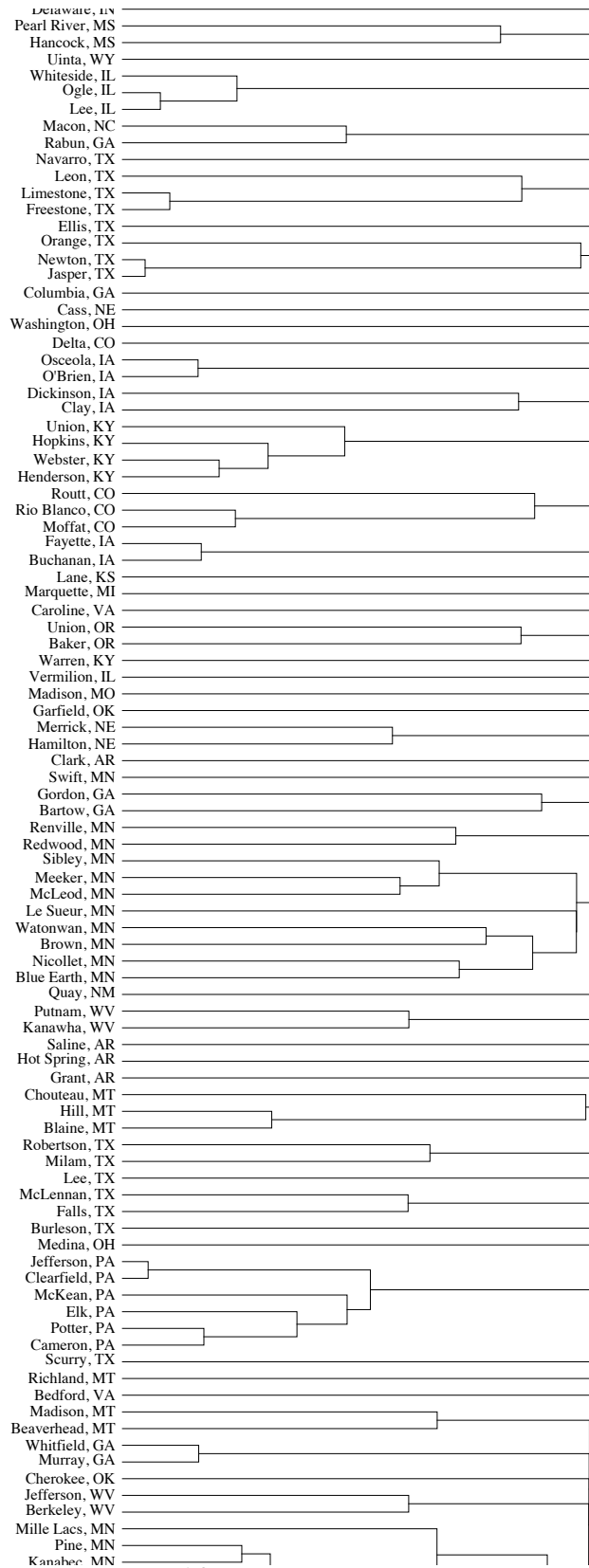


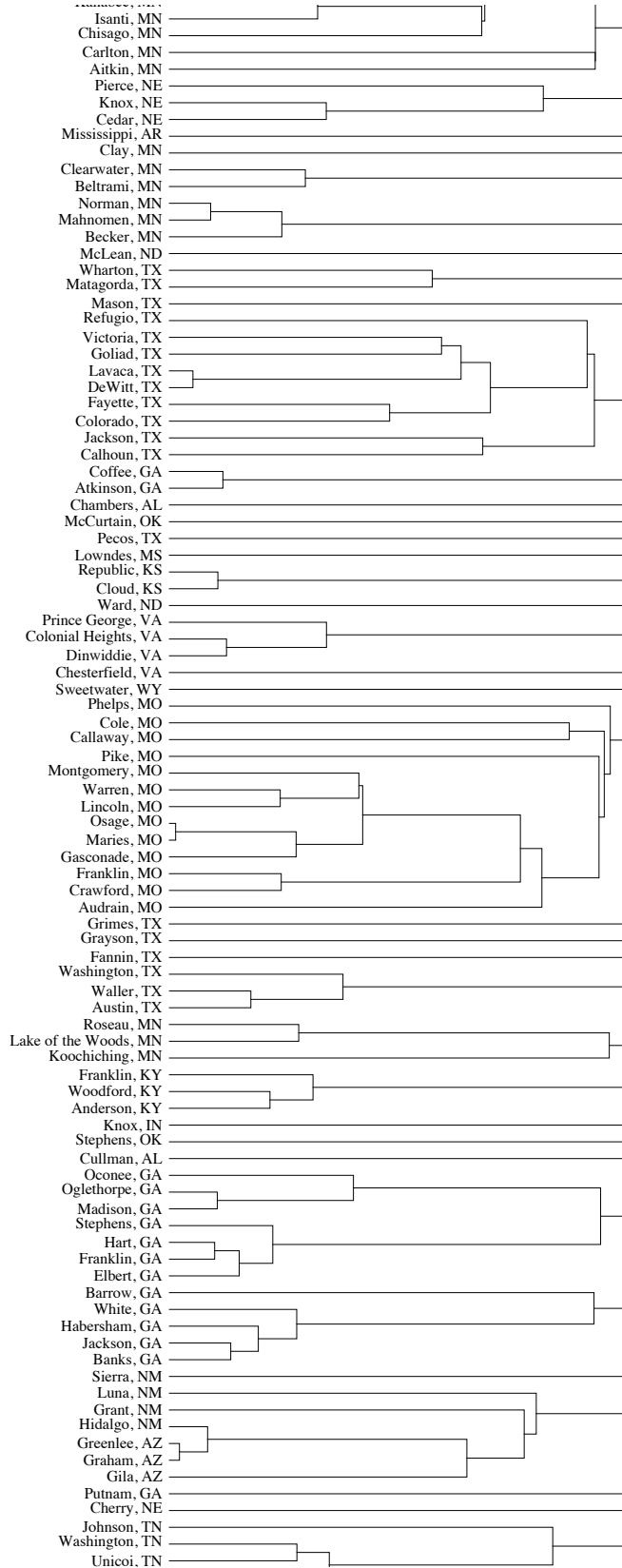


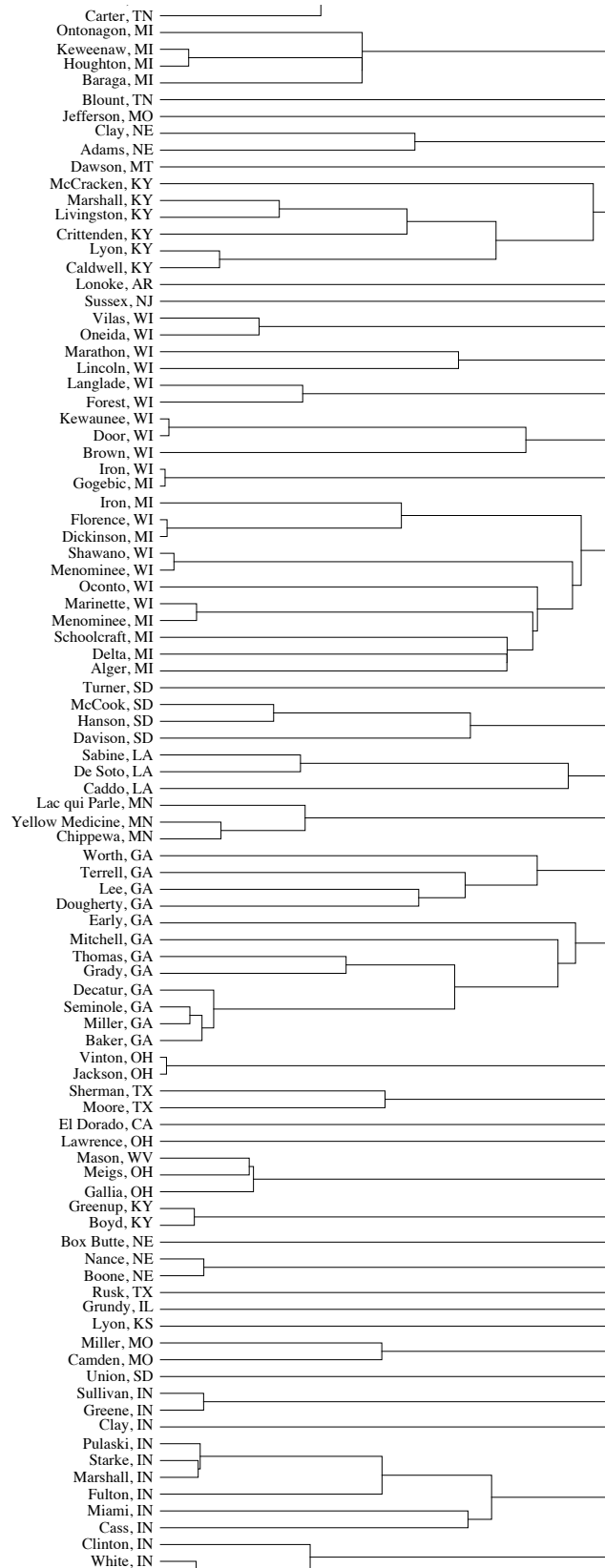


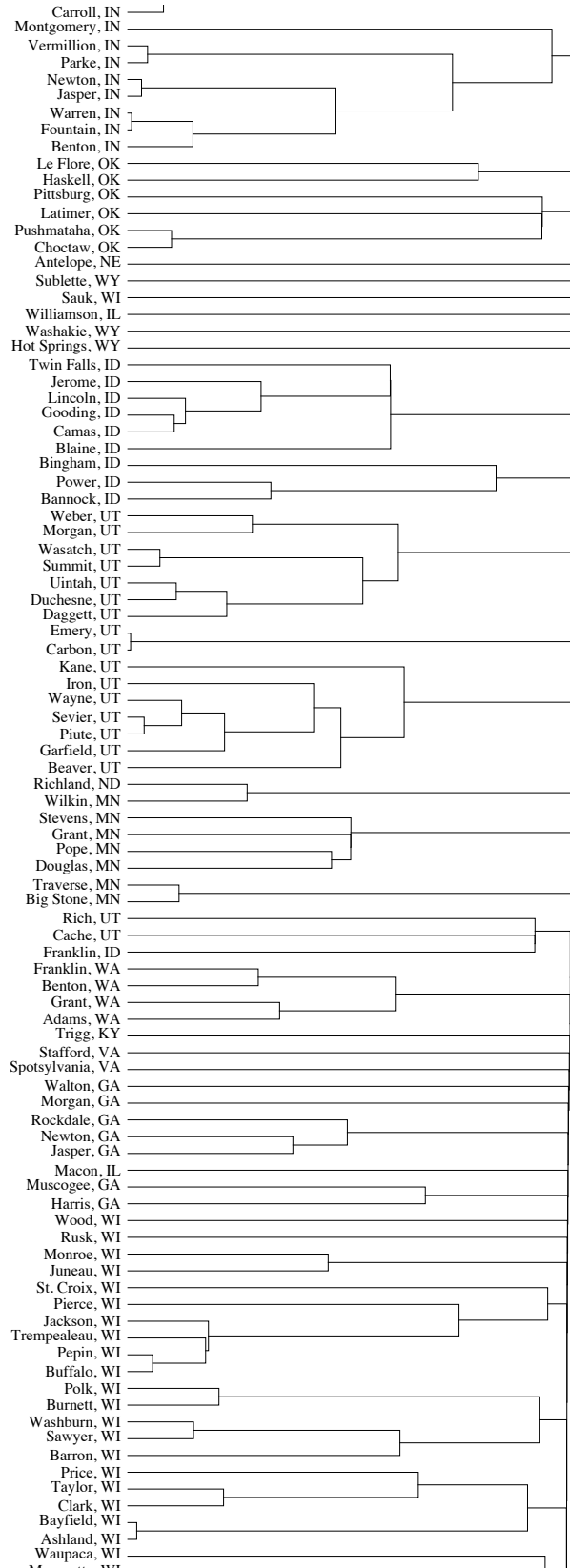


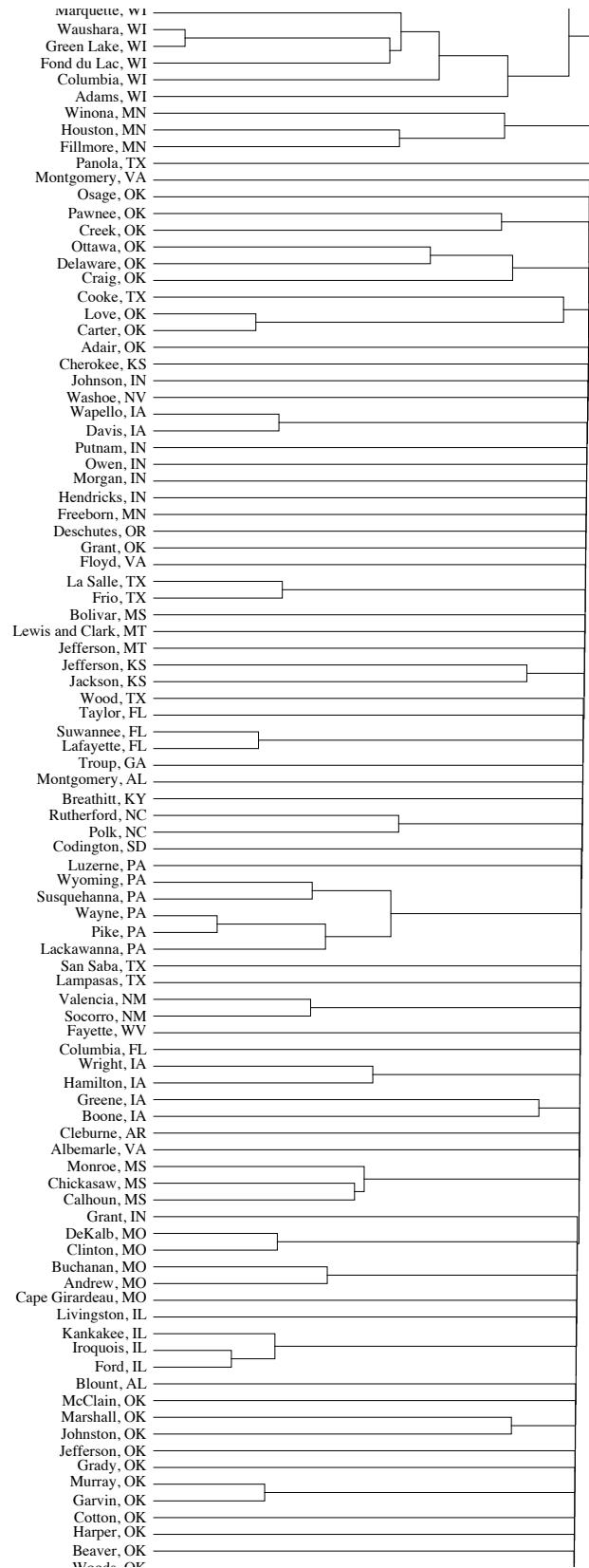


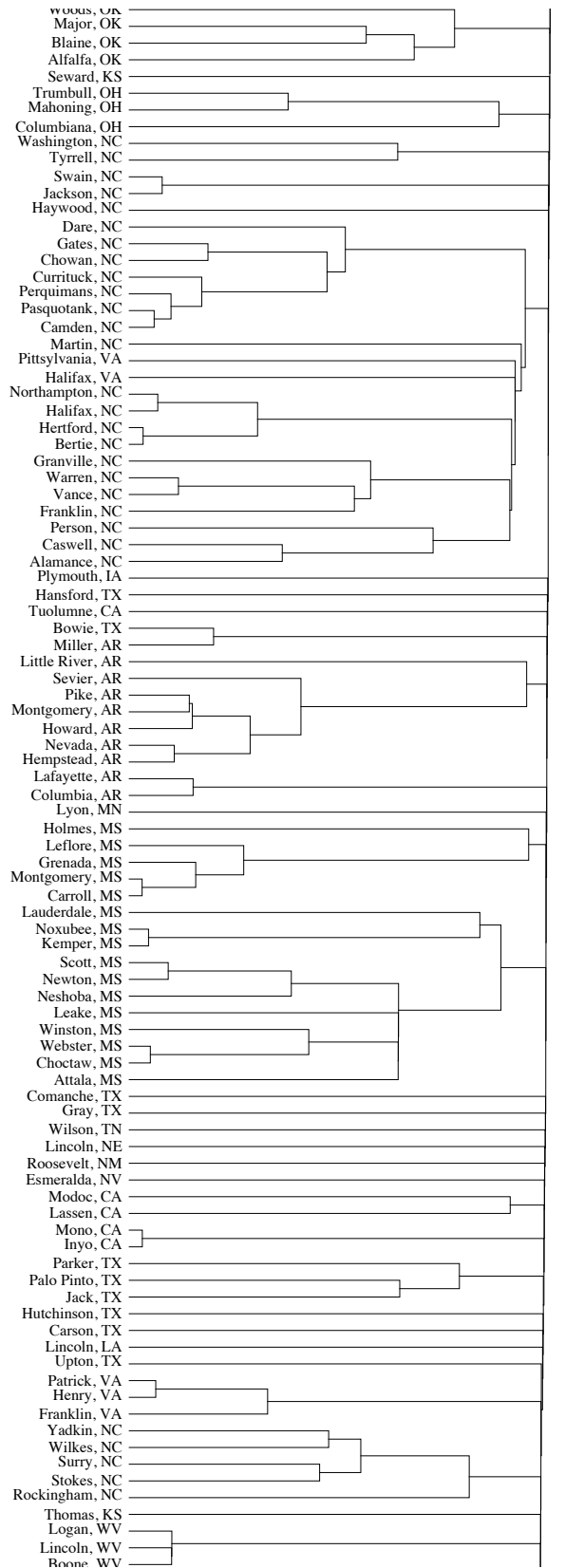


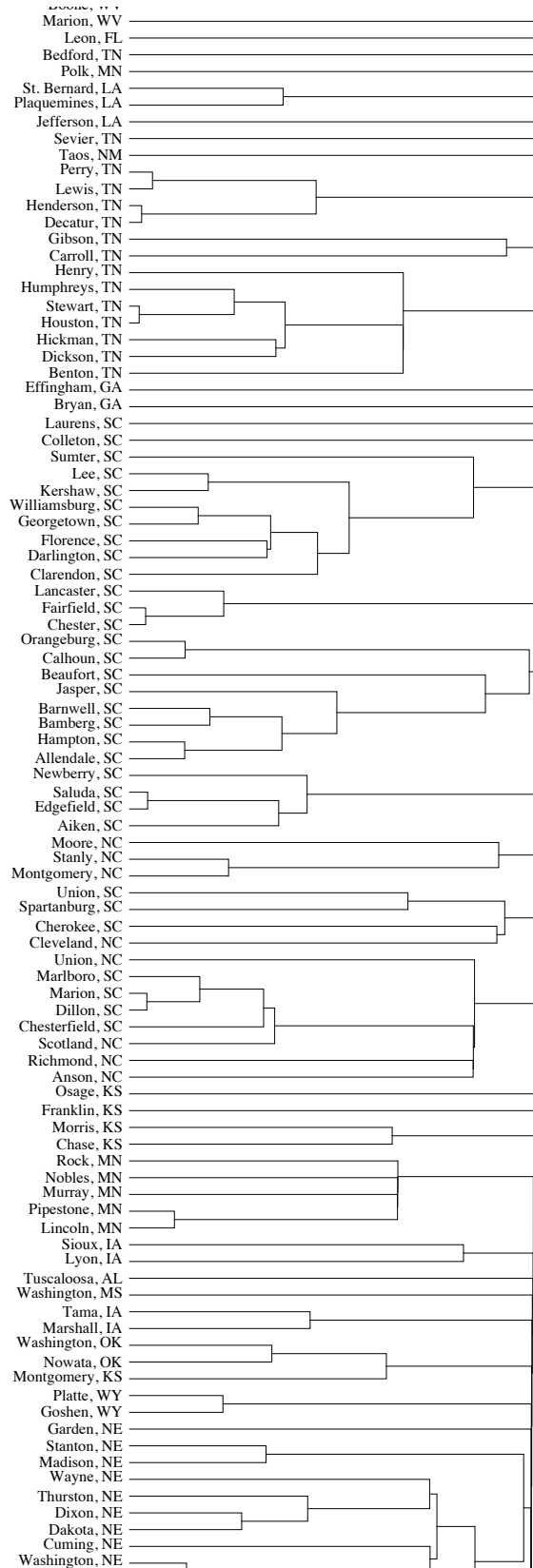


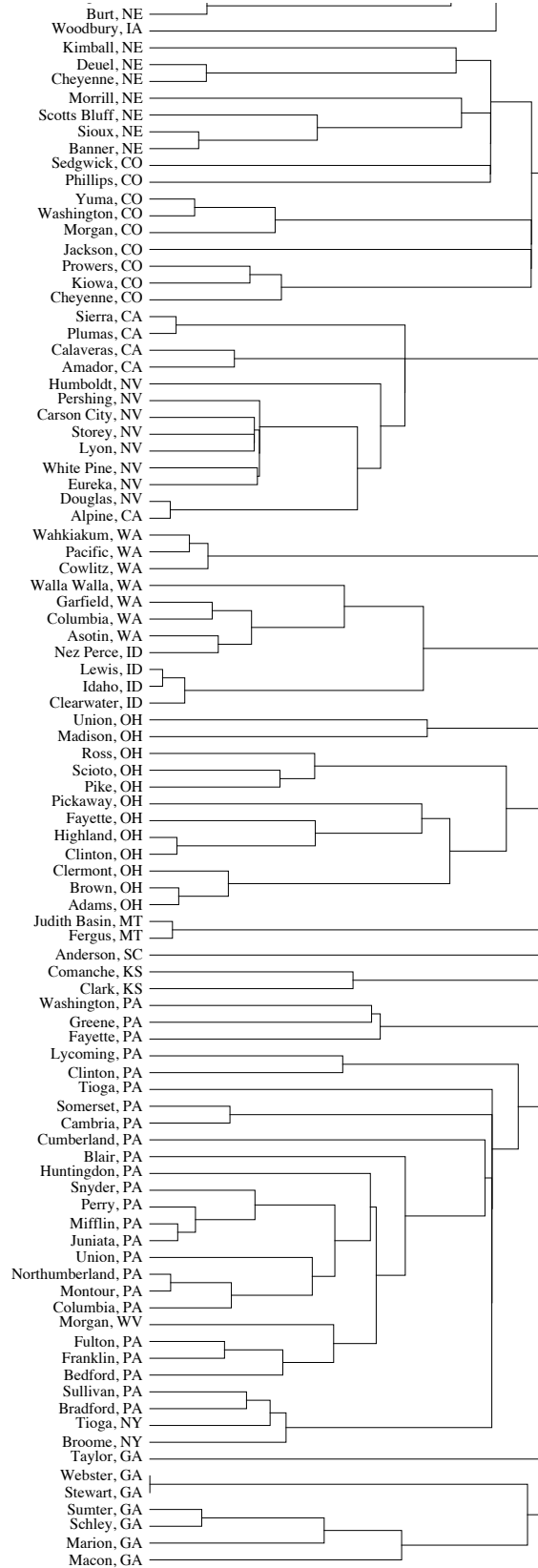


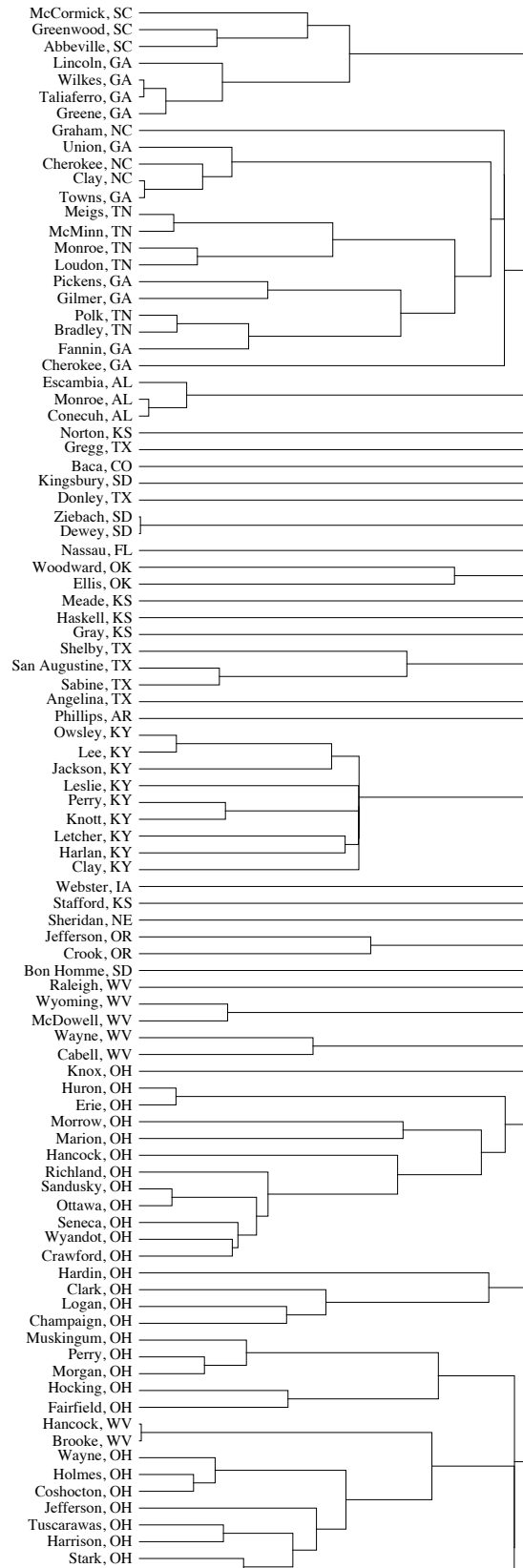


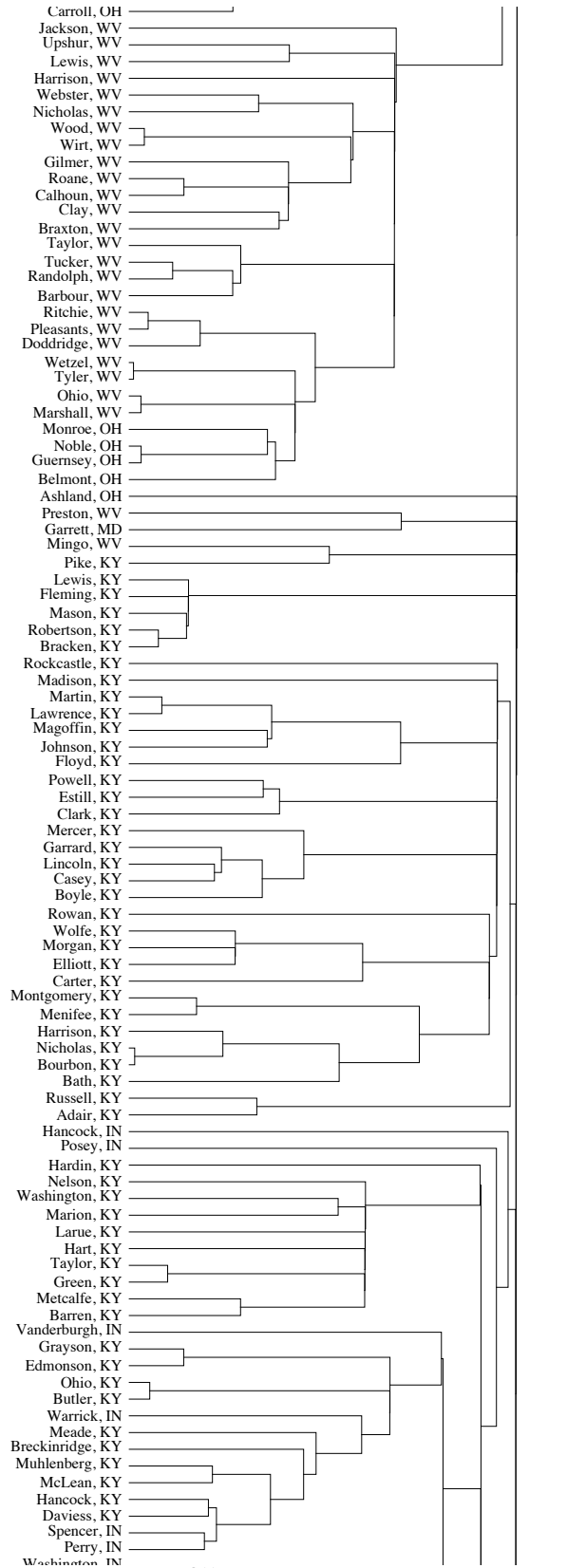


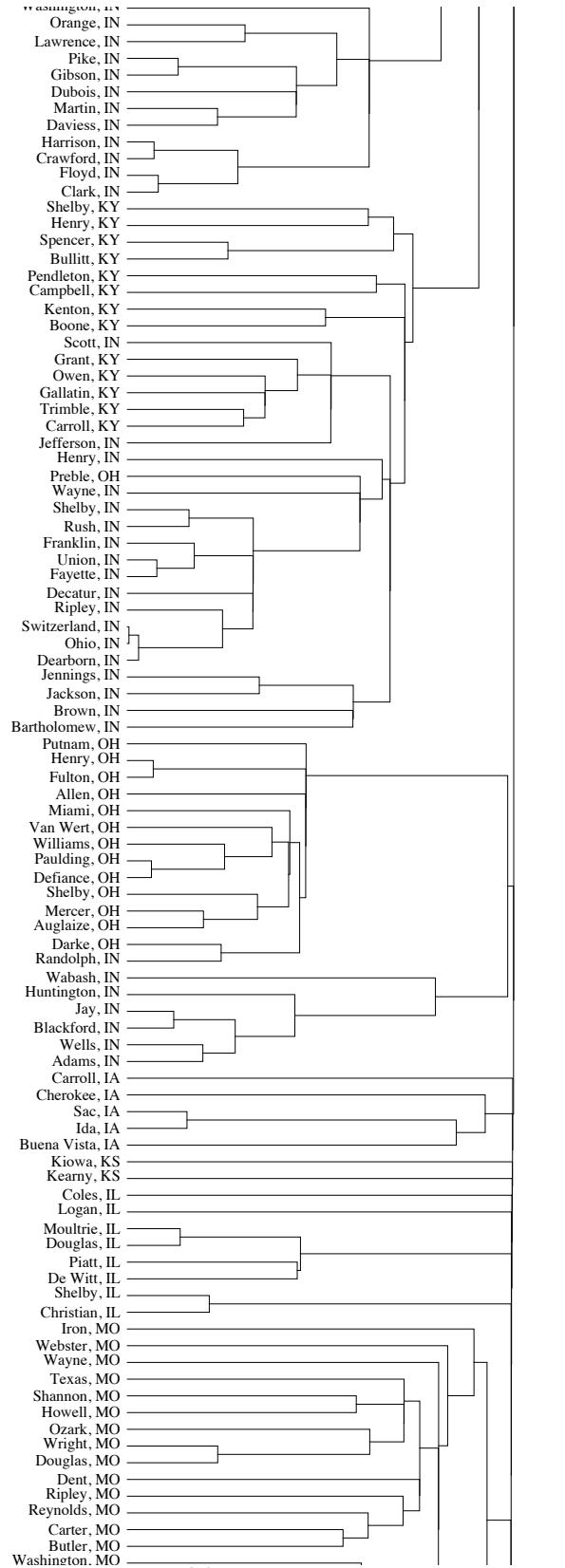


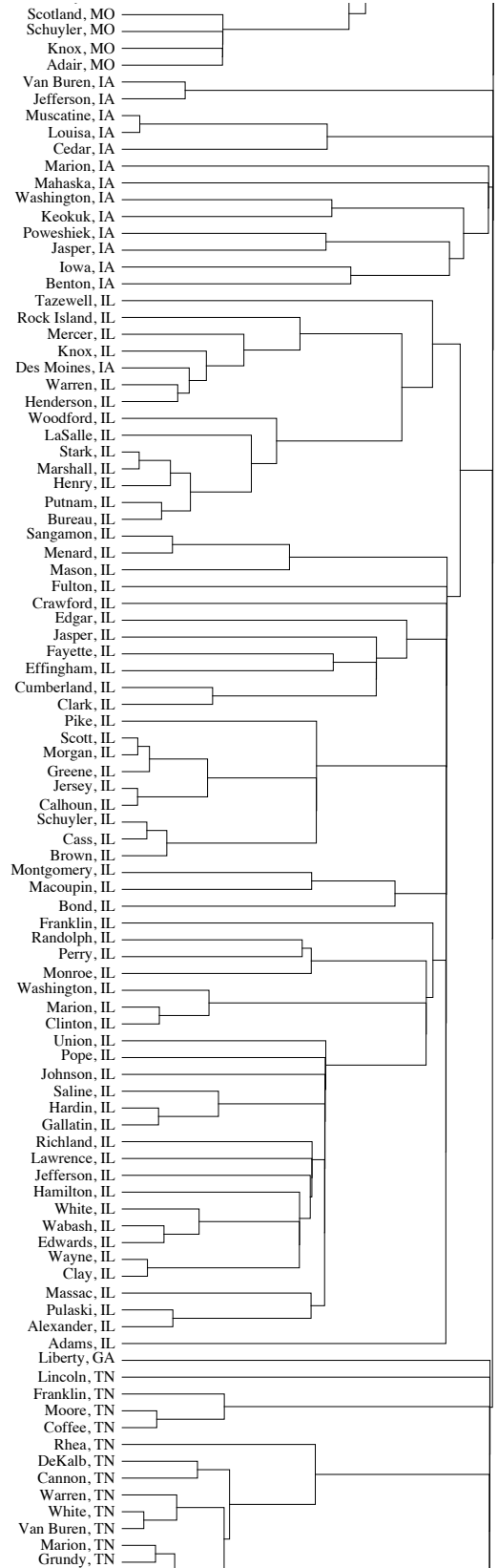


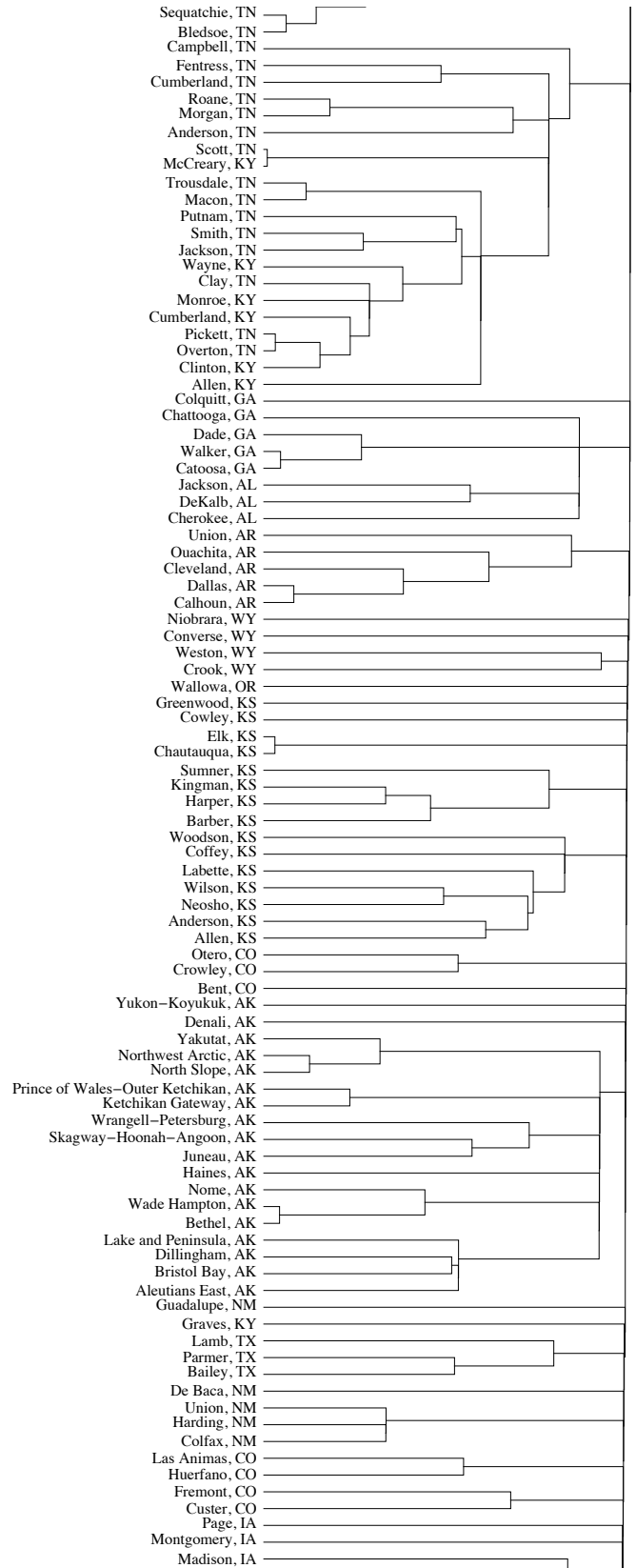


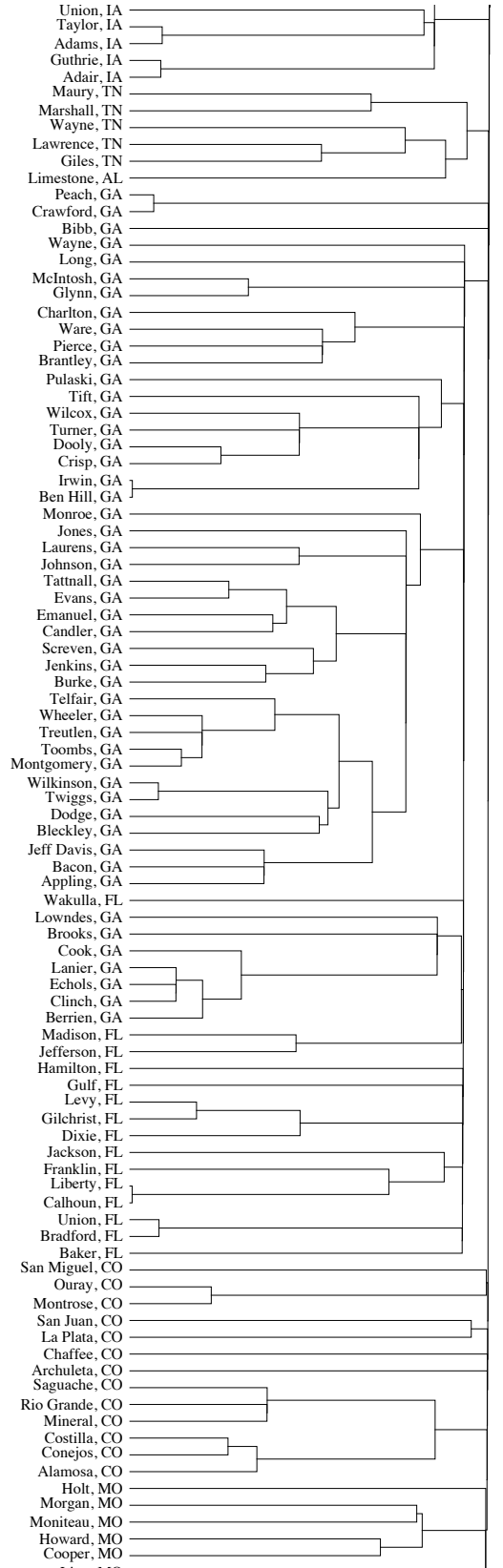


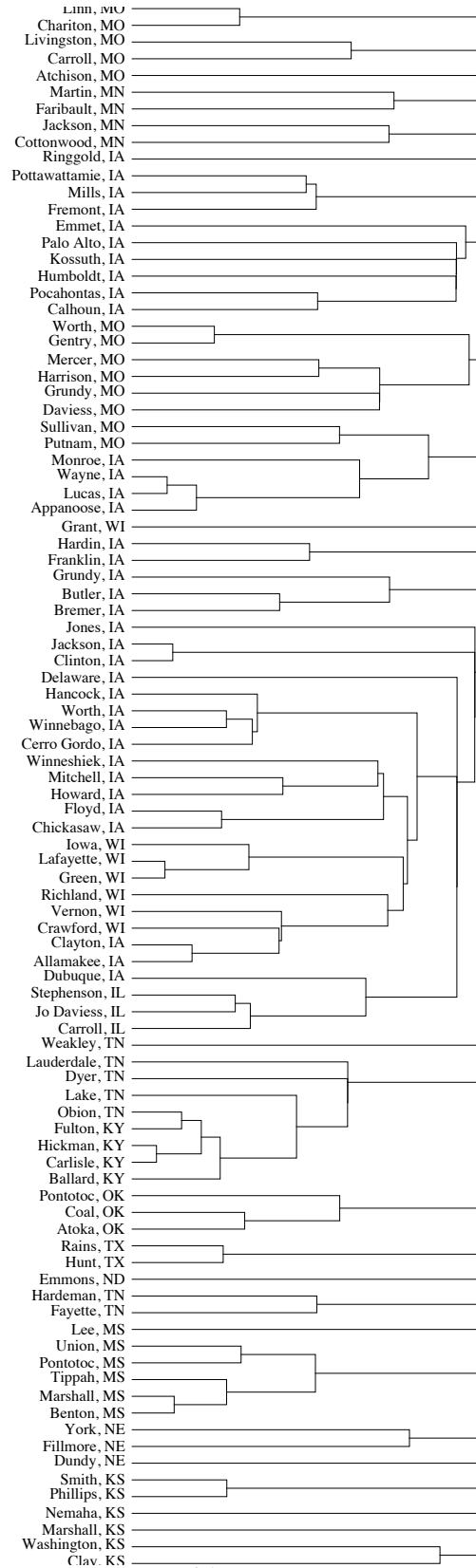


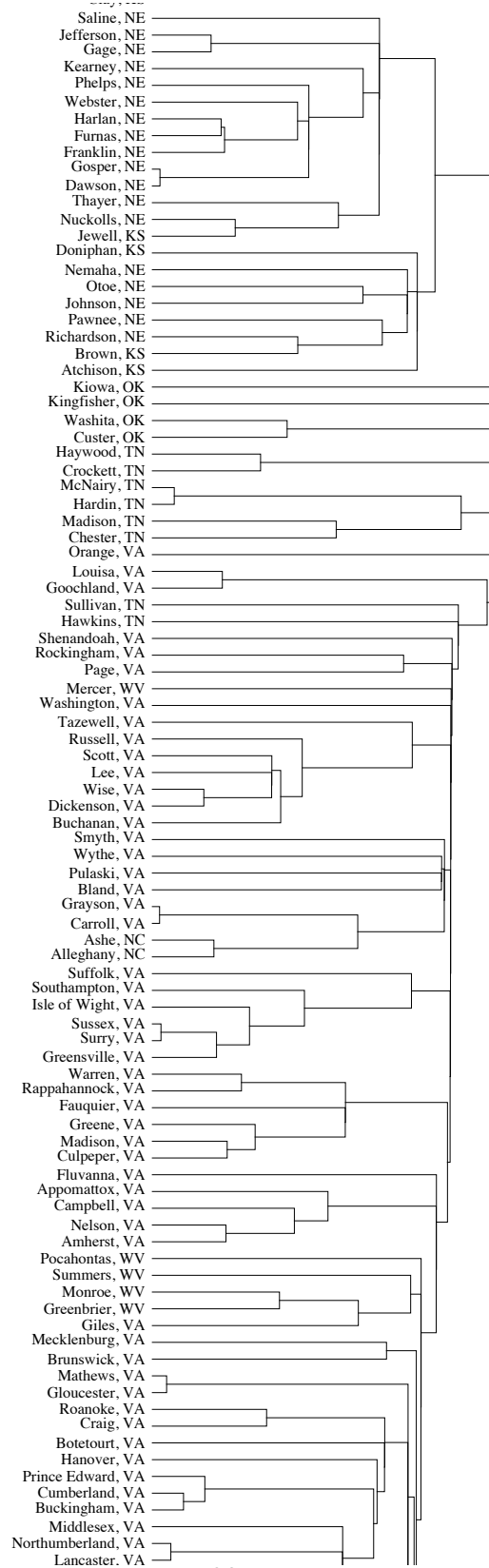


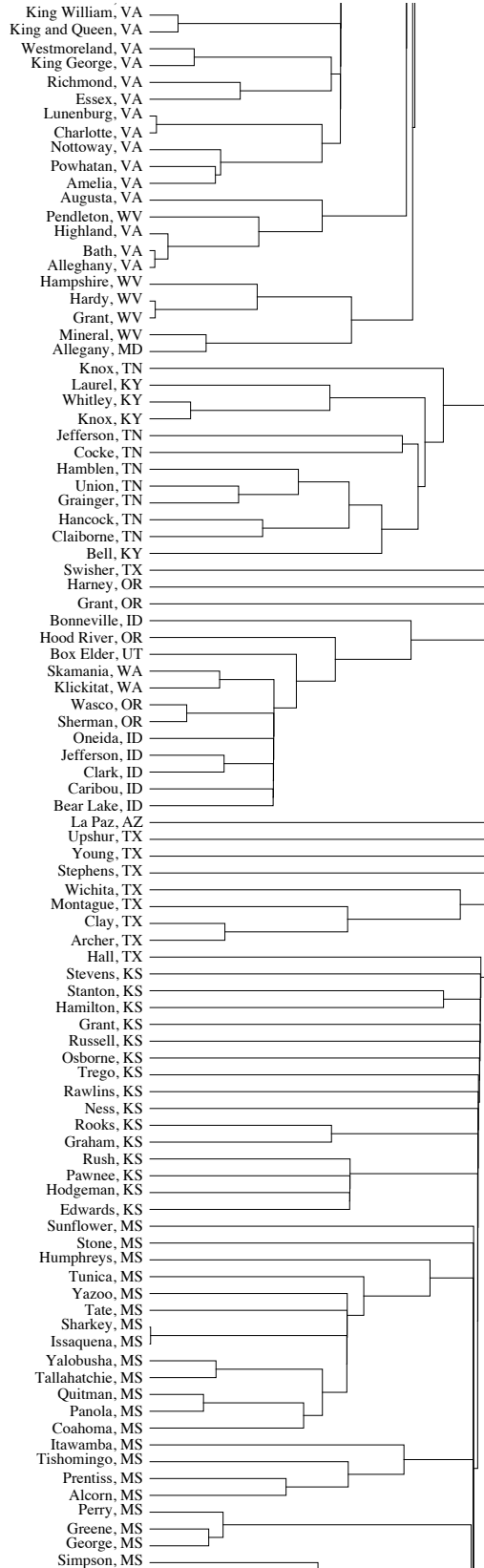


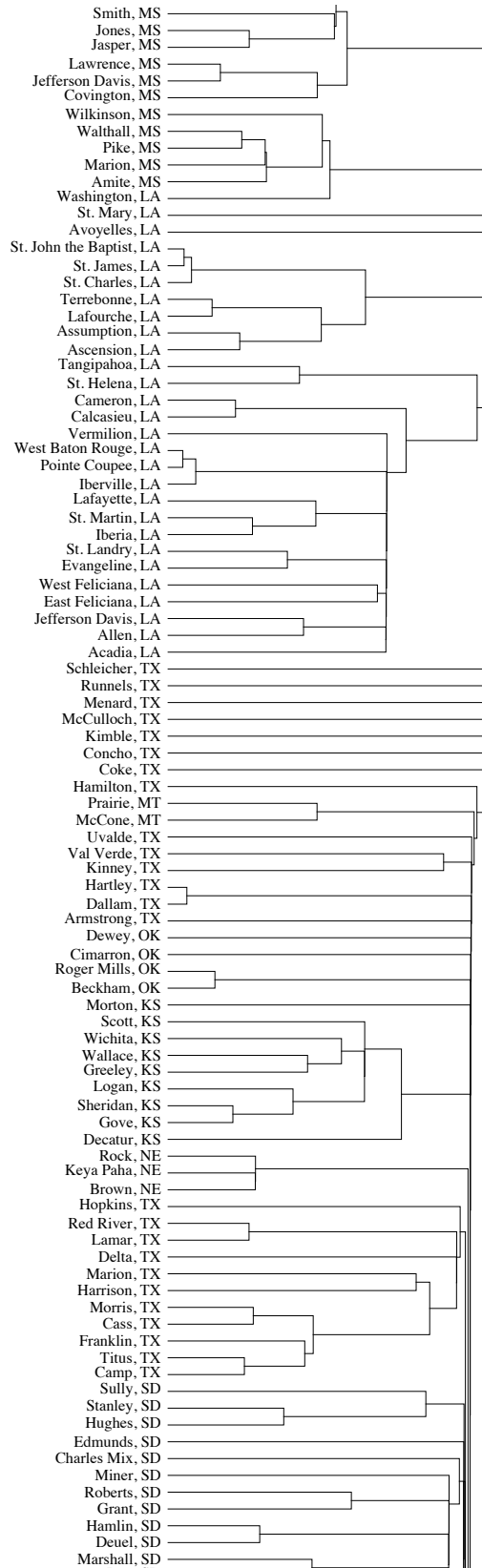


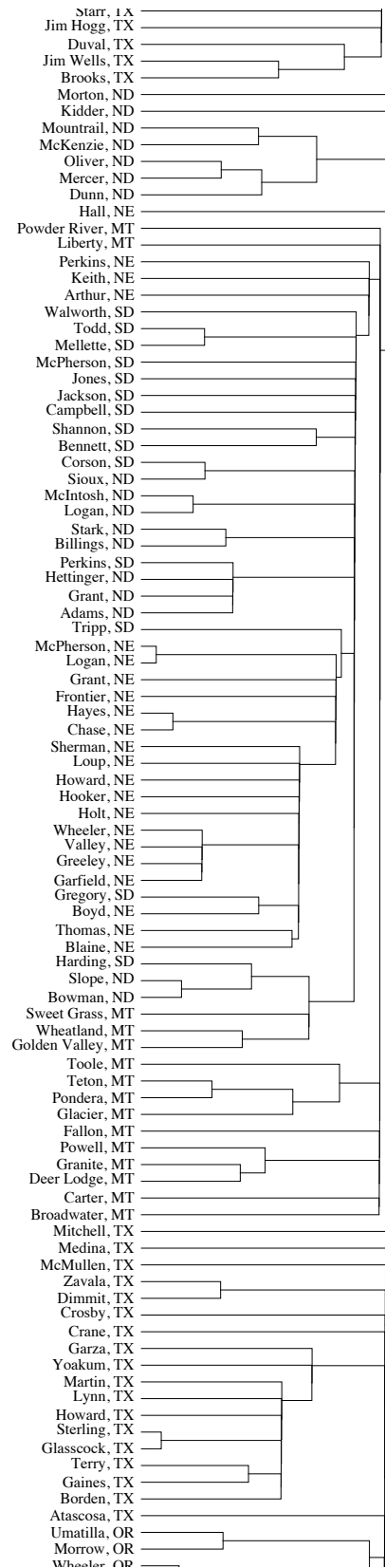


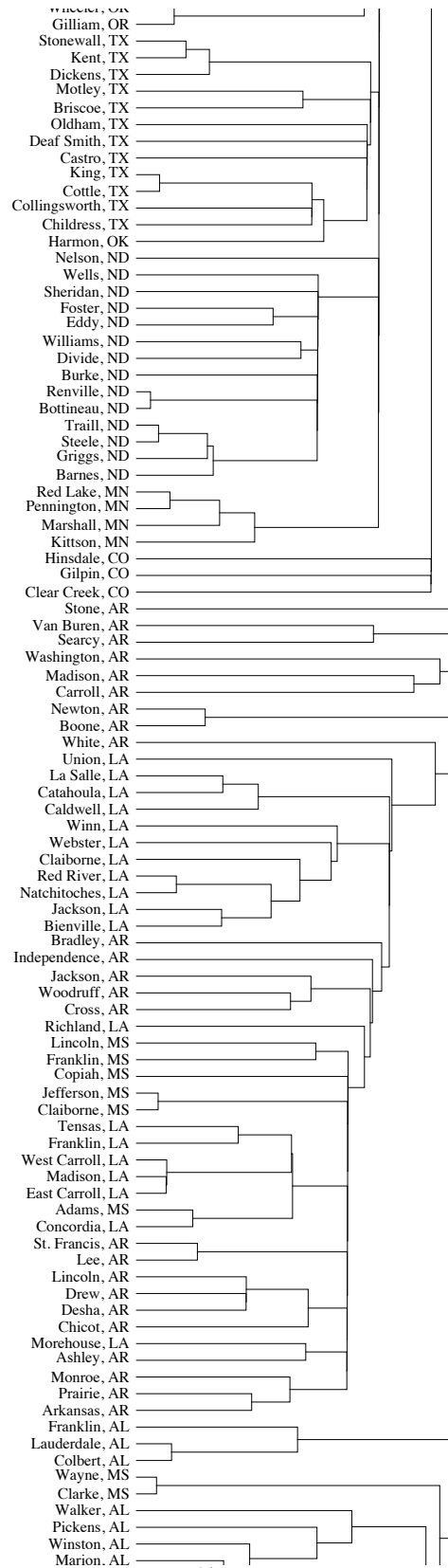


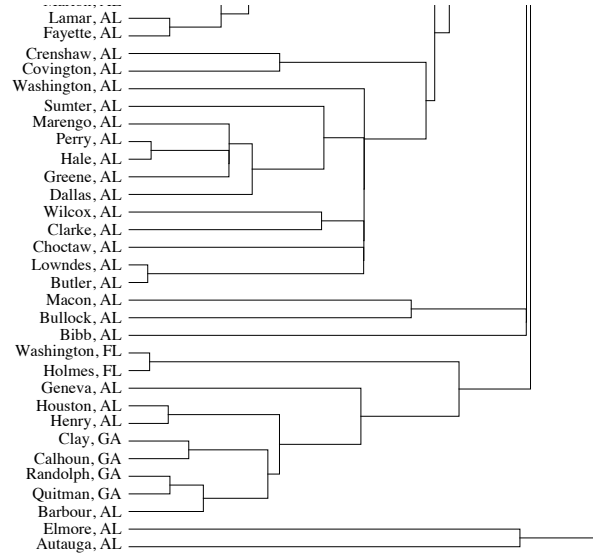












Acknowledgments

I would like to express appreciation to the Kavli Institute for Theoretical Physics (KITP) for computational support in this research.

- [1] P. B. Slater, *Comparative bi-stochastizations and associated clustering/regionalizations of the U. S. intercounty migration network*, arXiv:1208.3428.
- [2] C. Févotte and J. Idier, *Neural Computation* **23**, 2456 (2011).
- [3] P. B. Slater, *Proc. Natl. Acad. Sci.* **106**, E66 (2009).
- [4] R. C. Dubes, *J. Classif.* **2**, 141 (1985).
- [5] F. Mosteller, *J. Amer. Statist. Assoc.* **63**, 1 (1968).
- [6] R. E. Tarjan, *Info. Proc. Lett.* **17**, 37 (1983).
- [7] P. A. Knight, *SIAM. J. Matrix Anal. Appl.* **30**, 261 (2008).
- [8] F. Wang, P. Li, and A. C. König, in *IEEE International Conference on Data Mining* (IEEE Computer Society, 2010), pp. 551–560.
- [9] F. Wang, P. Li, A. C. König, and M. Wan, *Knowl. Inf. Syst.* **32**, 351 (2012).
- [10] P. B. Slater, *Dendrogram/regionalization of U. S. counties based upon migration flows*, arXiv:1207.0437.
- [11] I. S. Dhillon and J. A. Tropp, *SIAM J. Matrix Anal. Appl.* **29**, 1120 (2007).